JFMSE, 29(3), pp. 757~766, 2017.
수산해양교육연구, 제29권 제3호, 통권87호, 2017.
www.ksfme.or.kr
http://dx.doi.org/10.13000/JFMSE.2017.29.3.757

초등학생의 성별에 따른 차별기능문항 분석: 수학 과학 성취도 국제비교연구(TIMSS) 2007 수학영역을 중심으로

이승배 · 김석우†
(부산대학교)

Detecting Differential Item Functioning based on Gender: Field of Mathematics in the TIMSS 2007

Seungbae LEE · Sukwoo KIM†
(Pusan National University)

Abstract

This study investigated not only the existence of differently functioned item due to gender but also domain. In this study, the randomly selected data of TIMSS 2007, which consist of 681 male and 646 women, were analyzed.

To detect differently functioned items, this study employed Raju method. For Raju method, three-parameter logistic model was selected. Signed and unsigned area between two item characteristic curve were measured within the real ability range. An item which was detected commonly SA and UA area in Raju method was defined as a differently functioned item.

As a result of this study, six items among twenty seven items of mathematics in the TIMSS 2007 were differently functioned item. Five items among those six items, were in favor of boys and one item was in favor of girls. Number, Geometric Shapes and Measures, and Applying were in favor of boys. but Data Display, Reasoning were in favor of girls.

The conclusion of this study was summarized as existing differently functioned items in TIMSS 2007 and difference between favorable domain based gender. Finally, it is desirable to consider the differently functioned items by relating those item content for improving the test reliability of TIMSS 2007.

Key words : DIF, TIMSS 2007, Gender, Elementary school

Ⅰ. 서 론

1. 연구의 필요성 및 목적

일반적으로 남학생이 여학생보다 수학, 과학을 잘하고 여학생이 남학생보다 국어, 영어를 잘 한다는 통념이 있다. 성별에 따른 영역별 성취도의 차이가 나타나는 이유에 대한 연구는 다양하게 이루어져 왔다. 우선 교과별 성취도에 있어서 성차에 대한 생물학적 접근은 남성이 논리적인 연산과 관련된 좌뇌가 발달하고 여성이 언어와 관련된 우뇌가 발달하여 여성이 남성에 비하여 언어와 외국어 영역에서 높은 성취도를 보인다는 것이다. 또 다른 관점에서 수업 시간에 교사로부터 남학생과 여학생이 다른 대우를 받기 때문이

† Corresponding author : 051-510-2628, swkim@pusan.ac.kr
※ 이 논문은 2013년 이승배의 석사학위 논문을 수정·보완한 것임.

라는 주장도 있다. 또한 시험 문제가 어느 한 집단에게 유리하게 출제되었다는 주장이 있다. 여학생이 남학생에 비해서 외국어 영역에 대하여 수월성을 보이는 이유에 대해서 다양한 측면에서 연구되어질 수 있지만, 먼저 외국어 능력을 측정하는 도구를 살펴볼 필요가 있다. 어떠한 형태의 측정이던지 타당도, 신뢰도, 난이도, 변별도 등 일정한 측정학적인 준거들이 충족되어야 한다(Hambleton, 1989).

검사가 측정하고자 하는 바를 어느 특정 집단으로 치우침 없이 제대로 측정하고 있는가를 확인하는 것은 편파성 연구로, 이는 문항이 남성과 여성 집단 중 어느 특정집단에게 특히 어렵거나 쉽게 여겨진다면 그 문항은 편파적인 문항 또는 차별적으로 기능하는 문항이라고 할 수 있다. 차별적으로 기능하는 문항이란 인간의 잠재적 특성을 측정하는 검사 문항들이 능력이 같음에도 불구하고 집단의 특성 때문에 어떤 집단에서는 쉬운 문항으로 다른 집단에서는 어려운 문항으로 기능하는 것을 말한다(Kim Sukwoo, 1991; Kim Sukwoo·Park Haejin, 1994). 차별적으로 기능하는 문항은 검사의 측정학적 특성상의 문제일 수도 있지만, 교육기회의 균등 또는 문화적 환경과 보충학습 기회의 제공 등과 같은 문제 때문일 수도 있다(Linn & Harnisch, 1981; Doolittle, 1984; Lehman, 1986; Doolittle & Cleary, 1987; Muthen, Kao & Burstein, 1988; Kim Sukwoo, 1991; Kim Sukwoo & Park Haejin, 1994; Kim Shin-young, 2001). 따라서 성차에 관한 연구에서 성별 집단에 따라 문항이 차별적으로 기능하는지 아닌지를 파악하는 것은 매우 중요하다. 집단 또는 개인의 능력을 비교하기 위해서는 검사가 어떤 집단에서나 공정하다는 것을 전제로 하여야만 가능한 것이기 때문에 차별기능문항 연구는 활발하게 이루어져야한다.

평가문항이 집단별로 같은 기능을 수행하고 있는가의 여부는 평가도구의 공정성(fairness)이라는 문제로 대규모 평가나 고부담 평가 상황에서 오랫동안 논의되어져 왔다(Harnish, 1994). 만약 평가 문항이 공정성을 해친다면 측정학적 방법론은 공정성을 해치는 문항을 찾고 제거하기 위한 적절한 방법을 모색해야 한다. 측정학적 방법론에 바탕을 둔 편파성 문항 제거는 대단위 대규모 검사를 제작할 때 검사 도구를 검증하는 단계에서 이루어지는 것으로서 차별기능 문항의 연구는 단순히 측정학적 방법론에 의해 검사의 타당도를 위해 그 문항을 제거하기 위한 것이 아니라 편파성의 원인이 문항의 형식과 내용 중 어디에서 기인했는지 찾고 그 결과를 추후 분석이나 추후 평가에 반영하여 검사의 공정성을 더 높이는 노력의 일부분이 된다.

차별기능문항에 대한 연구는 1970년대부터 지금까지 활발히 이루어지고 있으나 우리나라에서는 차별기능에 대한 관심이 1990년대 후반부터 시작하여 Ahn Chang-kyu와 Cha Kyung-Ok(1984), Kim Sukwoo(1991), Seong Tae-je(1992) 등이 우리나라에 차별기능 문항 이론을 소개하였고, 이후 Choo, Jeong-A와 Seong Tae-je(1993), Seong Tae-je(1994)가 대학수학능력시험 실험평가의 차별기능문항을 추출하는 연구를, Kim Sukwoo와 Park Haejin(1994)이 성별 및 전공계열에 따른 차별기능문항에 관한 연구를 수행하였다. 그 후 차별기능문항 추출 방법을 비교하는 연구(Seong Tae-je, 1994; Song Miyoung, 2001)와 차별기능문항의 원인을 탐색하기 위한 연구(Kim Sukwoo, 1991; Kim Shin-young, 1993; Kim Sukwoo·Park Haejin, 1994; Chin Sujung & Seong Tae-je, 2004; Lee Myung-Ae, 2010)들이 수행되었다.

우리나라보다 앞서 차별기능문항에 관심을 가진 외국에서의 연구를 보면, Williams(1971)는 전통적인 고용시험과 교육시험이 백인 중류계층 문화에 유리하게 제작되어 있기 때문에 소수 집단의 진정한 능력을 반영하지 못한다고 주장하였으며 Faggen-Stecker, McCarthy, 그리고 Tittle(1974)도 표준화 검사 도구들이 남성명사를 여성명사보다 많이 사용하고 있어서 남성들에게 보다 유리

하게 작용한다고 하였다. 외국의 이러한 연구의 뒤를 이어 국내에서 고전검사이론과 함께 문항반응이론을 이용하여 편중문항을 밝히기 위한 연구와 통계적 방법들 간의 비교연구가 이루어진 것은 비교적 최근의 일이다(Kim Sukwoo, 1991; Kim Shin-young, 1993; Seong Tae-je, 1993; Choo, Jeong-A, 1993; Song Miyoung, 2001; Lee. Young, 2002; Noh Un-Kyung, 2007; Lee Myung-Ae, 2010).

본 연구는 TIMSS 2007 수학영역(4학년)에서 남학생과 여학생 집단에게 유리 또는 불리한 차별기능문항을 추출하고 추출된 차별기능문항이 측정하고자 하는 내용영역 및 인지영역에 대한 성별에 따른 차이 분석을 목적으로 한다. TIMSS는 4년마다 시행되는 국제 비교연구로 가장 최근에 시행된 것은 2015년이다. 그러나 본 연구에서 TIMSS 2007의 문항을 사용한 것은 2007년 당시 우리나라의 초등학교 4학년은 국제비교연구에 참여하지 않아 TIMSS 2007 검사도구에 대한 검증이 이루어지지 않아 전체 문항의 27개 문항을 추출하여 사용하였다. 차별기능을 추출하는 방법은 문항반응이론에 근거한 차별기능문항 추출방법 중 가장 타당한 Raju(1988)의 문항특성곡선(item characteristic curve) 간의 면적 계산방법을 사용하고 문항이 측정하고자 하는 내용영역 및 피험자의 인지영역은 국제 교육성취도 평가 협회(International Association for the Evaluation of Educational Achievement)의 분류를 따른다.

Ⅱ. 이론적 배경

1. 문항반응이론

가. 문항반응이론

검사도구를 구성하는 문항들의 질(quality)을 평가·관리하기 위하여 피험자가 응답한 자료를 가지고 문항에 대한 정보를 얻을 수 있는 이론을 검사이론이라고 한다. 검사이론은 크게 고전검사이론(Classical Test Theory)과 문항반응이론(Item Response Theory)으로 나뉜다.

고전검사이론은 검사 점수에 의해서 검사나 문항을 분석하는 방법으로 그 적용과 해석이 용이하여 현재 널리 사용되고 있다. 하지만 검사도구의 총점에 의해 분석하기 때문에 검사의 난이도와 피험자 집단의 특성에 따라 분석결과가 달라지는 단점이 있다.

반면에 문항반응이론은 문항 하나하나의 독특한 특성을 지닌 문항특성곡선(Item Characteristic Curve)에 의해 문항을 분석하기 때문에 피험자 집단이 달라져도 문항모수와 능력모수가 불변하는 장점을 가지고 있다(Seong Tae-je, 2001).

나. 문항특성곡선(Item Characteristic Curve: ICC)

문항특성곡선이란 피험자의 능력 θ에 따른 문항의 답을 맞힐 확률을 나타내주는 곡선을 말한다. 문항특성곡선에서 X축을 나타내는 준거변수인 피험자의 능력은 θ로 표기하며, Y축은 피험자의 능력 수준에 따라 문항의 답을 맞힐 확률 $P(\theta)$를 나타낸다(Hambleton et al., 1991). 문항의 답을 맞힐 확률을 나타내는 곡선으로 각 검사문항에 대한 정답확률과 능력척도사이의 관계를 나타낸 것이다(교육학대백과사전, 1996). 문항특성곡선은 S자 형태를 가지고 있으며 X축은 인간의 잠재적 특성인 능력을 나타내며 이는 θ로 표기하고 범위는 무한대이다. Y축은 능력수준에 따른 문항의 정답확률을 나타내고 범위는 0에서 +1까지이다.

2. 차별기능문항

차별기능문항이란 같은 능력수준을 가진 피험자들이 그들이 속한 집단 특성 때문에 문항의 답을 맞힐 확률이 다르게 나타나는 편파성문항 즉, 집단에 따라 문항의 기능이 달라지는 것을 말한다(Hambleton et al., 1991).

편파성문항 연구의 궁극적인 목적은 참조집단(reference group)과 연구집단(focal group) 간의 차이를 일으키는 문항의 특성들을 확인 및 추출하

여 교육과정의 구성이나 학습지도 방안의 모색에 이바지하기 위한 것이라 할 수 있다(Tatsuoka et al., 1988). Kim Shin-young(1993)은 편파성을 야기시키는 요인들에 있어서의 각 집단 간의 특성을 파악할 수 있을 것이라 하였다.

편파성을 논의한 교육·심리측정학자들은 검사의 편파성과 함께 공정성(fairness)에 대해 논의하면서, Angoff(1982)는 편파성문항은 피험자의 반응에 근거해서 판단되어야 하고 공정성은 검사의 목적에 근거해서 판단되어야 한다고 하였다.

차별기능문항을 추출하는 방법은 Clearly와 Hilton(1968), Angoff와 Ford(1973)의 문항난이도를 변환하는 방법, Camilli(1979)의 χ^2방법, lord(1980)의 χ^2방법, Swaminathan과 Rogers(1990)의 로지스틱 회귀분석(Logistic Regression) 방법, Shealy와 Stout(1993)의 SIBTEST(simultaneous item bias test) 방법, 문항반응이론에 기초한 방법 등 여러 가지가 있다.

문항반응이론에 기초한 차별기능문항 추출 방법들은 두 집단 간 문항특성곡선을 비교하는 방법과 추정된 문항 모수치를 비교하는 방법, 자료에 대한 문항반응모형 간 적합도를 비교하는 세 가지 범주로 나눌 수 있다(Holweger & Weston, 1998). 세 가지 방법들 중 두 문항특성곡선의 비교, 즉 곡선 간 면적에 의하여 차별기능문항을 추출하는 방법이 보다 타당한 방법으로 인정받고 있다. 이는 문항반응이론에 의한 문항특성곡선 간 면적 추정 방법이 차별기능문항에 대한 보편적 정의와 일치하기 때문이다(Seong Tae-je, 1993). 두 집단에서 같은 능력을 가진 피험자가 어떤 문항에 대해 더 높거나 낮은 수행을 한다는 것은 문항반응이론에 의하면 어떤 문항의 문항특성곡선이 두 집단 간에서 다른 형태로 나타나는 것을 의미하게 된다.

문항반응이론에 근거한 차별기능문항 추출 방법들 중 두 문항특성곡선간의 넓이를 계산하는 방법은 Rudner(1977)에 의하여 최초로 도입되었다. 그러나 Rudner의 두 문항특성곡선간의 면적 측정방법은 능력척도를 여러 개의 같은 간격으로 세분하여 계산하므로 그 절차가 복잡하다는 단점이 있다. 반면, Raju(1988)는 두 문항특성곡선 간의 면적을 문항모수치를 이용하여 간단하게 계산하는 공식과 문항특성곡선 간의 면적에 대한 유의도 검증을 제안하였다.

두 문항특성곡선 간 면적을 추정하는 측정치는 두 가지로 나눌 수 있는데, 하나는 차별기능의 정도뿐 아니라 방향도 고려한 SA(Signed Area)이고, 다른 하나는 차별기능의 방향은 고려하지 않고 정도만 고려하는 UA(Unsigned Area)이며, Raju의 3모수 문항반응모형에서의 문항특성곡선 사이의 넓이를 추정하는 공식은 다음과 같다(Raju, 1988; Seong Tae-je, 1993).

$$SA = \int_{-\infty}^{\infty} (F_1 - F_2)\, d\theta$$

$$UA = \int_{-\infty}^{\infty} |F_1 - F_2|\, d\theta$$

$$F_1 = F_1(\theta) = c_1 + (1-c_1)[\frac{\exp(Da_1(\theta - b_1))}{1+\exp(Da_1(\theta - b_1))}]$$

$$F_2 = F_2(\theta) = c_2 + (1-c_2)[\frac{\exp(Da_2(\theta - b_2))}{1+\exp(Da_2(\theta - b_2))}]$$

$F_1(\theta)$와 $F_2(\theta)$는 3모수 문항반응모형 에 대한 문항의 답을 맞힐 확률을 나타내는 두 개의 문항특성곡선을 의미한다. a, b, c는 각각 문항변별도, 문항난이도, 문항추측도를 나타내는 문항모수치이며 θ는 능력모수치, D는 정규 오자이브 모형에 의한 로짓(logit)을 로지스틱 모형의 로짓과 일치시키기 위한 상수로서 1.702이다(Haley, 1952).

3모수 문항반응모형에서 문항추측도(c)가 다를 경우, SA는 $+\infty$, 혹은 $-\infty$가 되고 UA는 $+\infty$가 된다. 결국, 문항추측도가 다를 때 두 문항특성곡선 간의 넓이를 문항모수치에 의해 계산하는 것은 불가능하다(Raju, 1988).

Ⅲ. 연구 방법

1. 연구 자료

본 연구에서 사용한 자료는 TIMSS 2007 4학년 수학영역 70문항 중 임의로 27문항을 한국어로 번역하여 K광역시 11개 초등학교 4학년 1327명에게 실시하여 얻은 결과이다. 본 검사 결과의 차별기능문항을 추출하기 위하여 분석에 사용된 성별에 따른 피험자 수는 <Table 1>과 같다.

<Table 1> Subjects based on gender

classify	Frequency	proportion(%)
boys	681	51.3
girls	646	48.7
total	1,327	100.0

총 피험자 수는 1,327명이며, 남자가 681명으로 51.3%를, 여자가 646명으로 48.7%를 차지하였다.

사용된 문항은 측정하고자 하는 내용 영역에 대하여 수, 도형, 자료해석으로 분류되고, 인지 영역에 대하여 이해, 적용, 추론으로 분류되며 <Table 2>와 같다.

2. 분석절차

본 연구에서 분석한 TIMSS 2007 수학영역의 원자료는 EXCEL 컴퓨터 프로그램을 사용하여 입력한 파일을 사용하고, 기술통계치는 SPSS 21.0 프로그램을 사용하여 산출하였다.

BILOG-MG 컴퓨터프로그램으로 여자집단과 남자집단의 각각에서 문항과 능력모수치를 추정하였다.

BILOG-MG 컴퓨터 프로그램은 세 단계로 나누어 모수를 추정하는데, 첫째 단계(PHASE Ⅰ)는 입력자료를 컴퓨터 언어로 전환하여 둘째와 셋째 단계를 수행할 수 있도록 임시파일을 만들고, 고전검사이론에 의한 문항난이도와 변별도 지수를 산출한다.

<Table 2> Construction of item based on content and cognitive domain

classify		item no.	no. of items
content domain	number	1, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 20, 21, 22, 24, 27	16
	geometric shapes and measures	7, 8, 15, 16, 17, 23, 25	7
	data display	2, 18, 19, 26	4
cognitive domain	knowing	3,5,8,12,13,15,17,18,23,25,	10
	applying	4, 6, 7, 9, 10, 11, 14, 16, 20, 21, 22, 24, 26, 27	14
	reasoning	1, 2, 19	3

둘째 단계(PHASE Ⅱ)에서는 문항반응이론에 의하여 문항특성을 추정하는 방법의 하나인 주변최대우도추정법(Marginal Maximum Likelihood Estimation)으로 문항모수를 추정한다.

셋째 단계(PHASE Ⅲ)에서는 능력모수를 추정한다. 능력모수를 추정하는 방법으로는 최대우도(Maximum Likelihood Estimation: ML), 베이지안 최대사후추정(Bayesian Maximum a Posterior Estimation: MAP), 베이지안 기대사후추정(Bayesian Expected a Posterior Estimation: EAP)의 세 가지 방법이 있는데 본 연구에서는 베이지안 기대사후추정법을 선택하여 실행하였다.

차별기능문항 지수로서 두 문항특성곡선사이의 넓이를 추정하기 위해 EQUATE 컴퓨터 프로그램을 사용하여 두 집단에서 각각 산출한 문항 모수치를 동등화하였다.

차별기능문항을 추출하기 위해서는 문항추측도가 동일해야하지만 모든 문항에 대해 추측도가 같다고 가정하고 특정한 값을 문항추측도로 고정시킬 수 없다. 따라서 Kim과 Cohen(1991a)의 제

한된 능력범위에서의 면적 계산을 공식으로 사용하였다. 면적 계산을 위해 사용한 Kim과 Cohen(1991a)의 공식을 수학적 연산방식으로 컴퓨터 프로그램화한 IRTDIF 컴퓨터 프로그램을 사용하여 문항특성곡선의 면적을 계산하여 산출하였다.

3. 자료분석

본 연구에서 차별기능문항의 추출 근거로 사용한 차별기능문항 지수는 문항반응이론을 토대로 한 Raju 방법의 SA(Signed area)와 UA(Unsigned area) 지수이다.

문항특성곡선 간 면적 측정치에 의하여 차별기능문항을 추출하는 Raju 방법은 1-모수 모형, 2-모수 모형, 3-모수 모형, 그리고 추측 모수가 동일한 3-모수 문항반응모형에 대해서만 유의도 검증을 할 수 있다. 그러나 본 연구는 추측 모수가 다른 3-모수 문항반응모형을 적용하였기 때문에 두 문항특성곡선 사이의 면적 측정치에 대한 유의도 검증을 시행할 수 없다. 따라서 본 연구에서는 Linn et al.(1981)의 연구결과에서 사용한 SA 또는 UA 지수가 ±.3을 넘는 지수는 차별기능문항으로 판정하였다. 그리고 산출된 SA지수가 음수이면 연구집단으로 설정한 여학생집단에게 유리하고, 지수가 양수이면 참조집단으로 설정한 남학생집단에게 유리한 문항으로 판정하였다.

Ⅳ. 결 과

본 연구는 TIMSS 2007 수학영역에서 차별기능문항을 추출하기 위하여 정답 문항은 1점, 오답이거나 무응답인 문항은 0점으로 점수화 시켜 자료를 분석하였다.

문항 당 1점씩의 배점에 의한 TIMSS 2007 수학영역에서 성별에 따른 피험자 점수의 평균, 표준편차, 최고치, 최저치는 <Table 3>과 같다.

<Table 3> Descriptive statistics of total score

classify		M	S.D	max.	min.
gender	boys	19.76	8.497	27	0
	girls	19.03	8.076	27	6

TIMSS 2007 수학영역에서 모든 문항에 대한 문항차별기능 지수인 SA와 UA 지수를 계산하였고 결과는 <Table 4>와 같다.

<Table 4> Index of SA and UA based on gender

item no.	gender	
	SA	UA
1	-0.15086431	0.07093274
2	-0.36260109	0.31980947
3	0.17233437	0.24828075
4	-0.49517297	0.02221573
5	-0.29319302	0.25443779
6	-0.18199333	0.01539136
7	0.13599402	0.16559210
8	0.41985992	0.42235935
9	-0.10777401	0.21096181
10	-0.20257803	0.01255810
11	-0.05188967	0.10538286
12	0.33665604	0.34135721
13	0.03415175	0.09444258
14	-0.21983289	0.26331665
15	-0.02403268	0.16071590
16	-0.11692789	0.19418517
17	0.14300334	0.18041218
18	0.12384623	0.14397031
19	0.09281684	0.09856524
20	-0.26871857	0.12873471
21	0.07670763	0.15515671
22	0.34962612	0.34960812
23	0.30991925	0.32083843
24	0.39399266	0.39399266
25	-0.24807801	0.25232753
26	-0.15481230	0.01497535
27	0.24756541	0.25645979

문항특성곡선 사이의 면적지수인 SA와 UA 지수가 ±.3을 초과하면 차별기능문항으로 판정한다고 하였는데 <Table 4>에 따르면 TIMSS 2007 수학영역에서 성별에 따른 차별기능문항이 존재하는 것으로 나타났다. 추출된 차별기능문항 6문항 중 5문항은 남학생에게, 1문항은 여학생에게 유리한 것으로 나타났다. 측정하고자 한 내용영역에 따른 분류를 적용하면 수 4문항, 도형 1문항, 자료분석 1문항이며, 인지영역에 따른 분류를 적용하면 이해 3문항, 적용 2문항, 추론 1문항으로 나타났다. 남학생이 유리한 영역은 수, 도형, 적용영역이고, 여학생이 유리한 것으로 나타난 영역은 자료해석, 추론영역이다.

Ⅴ. 논의 및 결론

본 연구는 TIMSS 2007 수학영역(4학년)에서 남학생과 여학생 집단에게 유리 또는 불리한 차별기능문항을 추출하고 추출된 차별기능문항이 측정하고자 하는 내용영역 및 인지영역에 대한 성별에 따른 차이 분석을 목적으로 하였다. 영어로 제작된 전체 70문항 중 27문항을 한국어로 번역한 도구를 사용하여 얻은 결과이므로 본래의 검사를 정확하게 반영한다고 할 수는 없으나 국제비교연구에서 사용되는 검사를 사용했다는 점에서 본 연구에서 사용한 검사도구가 신뢰롭고 타당하다고 할 수 있으며 연구의 결과가 주는 시사점은 크다고 할 수 있다. 결과에서 보듯이 TIMSS 2007 4학년 수학영역에서 6문항의 차별기능문항이 추출되었다.

초등학교 학생 및 교사를 대상으로 성별 차이에 대한 인식을 조사한 Lee Dae Sik· Kim Su Mi(2003)에 의하면, 내용적 측면에서 도형, 확률과 통계영역은 남학생이 우수한 성취를 보이는 영역인 것으로 나타났으며, 본 연구의 결과와 일치한다. 하지만 수학과 국가수준 학업성취도 평가에서의 성별 차이를 분석한 Ko Jung Hwa, Do Jong Hoon, & Song Miyoung(2008)에 따른면, 도형영역에 대하여 여학생이 남학생에 비해 높은 성취를 보인다고 분석하였다. 이는 같은 내용영역 내에서도 각각의 문항이 갖고 있는 인지적 요소와 내용요소에 따라 성별 차이가 다르게 나타날 수 있음을 보여준다.

차별기능문항으로 추출된 문항들의 내용을 살펴본 결과, 2번 문항은 나무의 종류와 나무의 개수에 대한 표를 원그래프로 변환하는 내용이며, 8번 문항은 특정 도형을 90° 회전하였을 때의 모양을 예측하는 내용이다. 12번 문항은 '3 + 2십 + 4백' 과 같은 수를 찾는 내용이며, 22번 문항은 두 종류의 자전거 대여에 관한 안내문을 읽고 일정시간 이후의 대여료 크기를 비교하는 내용이다. 23번 문항은 종이를 잘라 모양을 변형한 이후의 넓이 차이에 대한 내용이며, 마지막 24번 문항은 상자의 색깔에 따라 다른 개수의 연필이 담긴 상자에 대하여 전체 연필 수를 추론하는 내용이다. 이와 같이 문항 자체가 남·여의 문화적 특성을 반영하는 언어로 구성되었다든지 특정한 성별의 성향을 나타낸다고 할 수 없다. 다만, 자전거와 관련된 22번 문항의 경우는 남학생에게 더 익숙한 자전거라는 소재로 인하여 남학생에게 더 유리한 문항일 것이라 추측할 수도 있다. 그러므로 본 연구에서 추출된 차별기능문항의 원인은 문항 자체의 결함보다는 남·여의 심리적인 특성의 차, 혹은 실제 교수·학습 과정이나 학습 기회에서의 차이 등을 의미하는 것일 수도 있다.

본 연구는 TIMSS 2007 수학영역에 대하여 성별에 따른 차별기능문항 추출과 추출된 차별기능문항이 측정하고자 하는 내용영역 및 인지영역에 대한 성별에 따른 차이 분석을 목적으로 하였으며 두 가지의 제한점을 지니고 있다.

첫째, TIMSS 2007 수학영역 70문항과 그 중 한국어로 번역하여 본 연구에서 사용한 27문항의 내용타당도에 대한 검증이 이루어지지 못하였다. 둘째, Raju의 방법을 사용하기 위해 Kim과 Cohen(1991a)의 제한된 능력범위에서의 넓이 계산 방법을 적용하였는데, IRTDIF(Kim & Cohen,

1991b) 컴퓨터 프로그램에서는 Kim과 Cohen의 공식(1991a)에 의한 SA(Signed Area)와 UA (Unsigned Area)에 대하여 유의도 검증이 이루어지지 않았다.

본 연구에 따른 후속연구를 제언하자면 통계적 방법에 의해 추출된 차별기능문항에 대하여 내용 전문가와 교육과정 전문가, 평가 전문가 등의 심층적인 내용 분석이 필요하며, 이러한 분석을 기초로 차별기능문항의 원인을 제거하여 특정 집단에 유리 또는 불리하지 않은 공정한 검사도구를 제작할 수 있는 기초자료를 제공할 필요가 있다.

References

Ahn, Chang-kyu & Cha, Kyung-Ok(1984). A Comparison of Four Statistical Procedures for Detecting Test Item-Bias. Social Survey Research, 3, 29~49.

Angoff, W. H.(1982) Use of item difficulty and discriminant indices for detecting item bias. ed. by Berk, R. A. In Handbook of method for detecting test bias, 96~116. Baltimore, London: The Johns Hopkins University Press,

Angoff, W. H. & Ford, S. F.(1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10(2), 95~106

Baker, F. B., & Al-Karni, A.(1991). EQUATE: Computer program for equating two metrics in item response theory [Computer program]. Madison WI: University of Wisconsin, Laboratory of Experimental Design.

Camilli, G.(1979). A critique of the chi-square method assessing item bias. Unpublished paper, Laboratory of Education Research, University of Colorado at Boulder.

Chin, Sujung & Seong, Tae-je(2004). An Exploratory Study on Item-format Related DIF with MH and SIBTEST techniques. Journal of Educational Evaluation, 17(2), 215~236.

Choo, Jeong-A & Seong, Tae-je(1993). Detecting gender differently functioned items of the 4th and 5th tryouts of College Scholastic Ability Test by Mantel-Haeszel and Raju Methods. Journal of Educational Evaluation, 6(2), 259~286.

Choo, Jeong-A.(1993). Detecting gender differently functioned items of the tryouts of College Scholastic Ability Test. Master's Thesis, Ehwa Womans University, Seoul.

Cleary, T. A. & Hilton, T. L.(1968). An Investigation of Item Bias. Educational and Psychological Measurement, 28, 61~75.

Cook, L. L. · Dorans, N. J. & Eignor, D. R.(1988). An assessment of the dimensionality of three SAT-Verbal test editions. Journal of Educational Statistics, 13, 19~43.

Doolittle, A. E.(1984). Interpretation of differential item performance accompanied by gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Doolittle, A. E. & Cleary, T. A.(1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24(2), 157~166

Educational Research Institute of Seoul National University(1998). The encyclopedia of education. Seoul: Hawoo.

Faggen-Steckler, J. · K. A. McCarthy, & C. K. Tittle (1974). A quantitative method for measuring sex bias in standardized test. Journal of Educational Measurement, 11, 151~161.

Greaud, V. A.(1987). Investigation of the unidimensionality assumption of item response theory. Master's Thesis, The Johns Hopkins University, Baltimore, Maryland.

Haley, D. C.(1952). Estimation of the dosage ,ortality relationship when the dose is subject to error(Technical Report No. 15). Stanford CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Hambleton, R. K. & Swaminathan, H.(1985). Item responsw theory : Principles and application. Boston: Kluwer Boston

Hambleton, R. K.(1989). Principles and selected applications of item response theory. In R. L. Linn (ed.), Educational measurement(3rd ed., pp. 147~200). New York: MacMillan

Hambleton, R. K. · Swaminathan, H. & Rogers, H. J.(1991). Fundmentals of item response theory. CA:

SAGE Publications.

Hambleton, R. K.・Swaminathan, H., & Rogers, J. H.(1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hanson, B. A.(1988). Uniform DIF and DIF defined by differences in item response function. Journal of Educational and Behavioral Statistics, 23(3), 244~253.

Harnish, D. L.(1994). Performance assessment in review : New direction for assessment student understanding. International Journal of Educational Research, 21(3), 341~350

Hatti, J. A.(1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49~78.

Holland, P. W. & Thayer, D. T.(1988) Differential item performance and the Mentel-Haenszel procedure. In H. Wainer & H, I, Braun(Eds.). Test validity(pp. 129~145). Hillsdale, NJ: Lawrence Erlbaum.

Holweger, N. & Weston, T.(1998). Differential item functioning: an applied comparison of the item characteristic curve method with the logistic discriminant function method. (ERIC Document Reproduction Service No. ED 422362)

Jensen, A. R.(1969). How much can we boost IQ and Scholistic achievement?. Harvard Educational Review, 39, 81~83.

Kim, S. H. & Cohen, A. S.(1991a). IRTDIF: A comparision of two area measures for detecting differential item function. Applied Psychological Measurement, 15, 269~278.

Kim, S. H. & Cohen, A. S.(1991b). IRTDIF: A computer program for IRT differential item functioning[Computer program]. Madison WI: University of Wisconsin.

Kim, Shin-young(1993). An Empirical Study for Validity of Mantel-Haenszel Method. Journal of Educational Evaluation, 6(1), 59~90.

Kim, Shin-young(2001). Methods of Analyzing Differently functioned items. Seoul: kyoyookkwahaksa.

Kim, Sukwoo.(1991). Gender and OTL effects on mathematics achievement for U.S. SIMS 12th grade students. Journal of Educational Evaluation, 4, 32~58.

Kim, Sukwoo. & Park, Haejin.(1994). A Study of Differential Item Functioning based on Gender and Major. Journal of Pedagogy in Pusan, 7, 45~64.

Ko, Jung Hwa・Do, Jong Hoon, & Song, Miyoung.(2008). An Analysis of the Gender Difference in National Assessment of Educational Achievement of Mathematics. The Journal of Educational Research in Mathematics, 18(2), 179~200.

Lee, Dae Sik & Kim, Su Mi.(2003). Elementary School Students` and Teachers` Responses on Sex Differences in Mathematics Learning. The Journal of Elementary Education, 16(1), 297~315.

Lee, Myung-Ae(2010). DIF Identification via Hierarchical Nonlinear Model. Journal of Educational Evaluation, 23(1), 171~190.

Lee. Young(2002). Detecting Differential Item Functioning based on Gender and Major : Verbal Area in the tryout of College Scholastic Ability Test. Master's Thesis, Pusan National University, Busan.

Lehman, J. D.(1986). Opportunity to learn and differential item functioning. Unpublished doctoral dissertation, University of Valifornia, Los Angeles.

Linn, R. L. & Harnish, D. L.(1981) Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109~118.

Lord, F. M.(1980). Applications of item response theory to practical testing problems. Hilldale, NJ: Lawrence Erlbaum.

Lord, F. M. & Novick, M. R.(1968). Statistic theories of mental test scores. Reading, MA: Addison-Wesley.

Mellenbergh, G. J.(1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105~108.

Mislevy, R. J. & Bock, R. D.(1990). BILOG 3 for windows: Item analysis and test scoring with binary logistic models[Computer program]. Mooresville IN: Scientific Software Inc.

Muthen, B.・Kao, C.-F. & Burstein, L.(1988). Instructional sensitivity in mathematics achievement test item: Application of a new IRT-based detection technique. Forthcoming in Journal of Educational Measurement.

Muthén, B. O. & Muthén, L. K.(2001). Mplus: Statistical analysis with latent variables[Computer program and manual]. Los Angeles: Statmodel.

Noh, Un-Kyung(2007). Detecting Gender-related Differential Item Functioning on the Spatial Ability Test in the Aptitude Battery for Middle School Student. Master's Thesis, Ehwa Womans University, Seoul.

Peterson, N. S.・Kolen, M. S. & Hoover, H. D.(1989). Scaling, norming and equating. In Educational Measurement(3rd). Ed by Linn, R. L.. New York: Macmillian Publishing Company.

Potenza, M. T. & Dorans, N. J.(1995). DIF Assessment for Polytomously Scored Item: A Framework for Classification and Evaluation. Applied Psychological Measurement, 19(1), 23~37.

Raju, N. S.(1988). The area between item characteristic curves. Psychometrika, 53, 495~502.

Roussos, L. A. & Stout, W. F.(1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 32(2), 215~230.

Runder, L. M.(1977). An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation. Catholic University of America, 18~27.

Scheuneman, J. D.(1975). A new method of assessing bias in test items. Paper presented at annual meeting of American Educatioal Research Association. Washington D. C..

Seong, Tae-je(1993). A Comparative Study between Raju and Mantel-Haeszel Methods for Detecting Differential Item Function. Journal of Educational Evaluation, 6(1), 91~120.

Shealy, R. T. & Stout, W. F.(1993). An Item Response Model for Test Bias and Differential Test Functioning. Hillsdale, NJ: Erlbaum.

Song, Miyoung(2001). Detection of gender-related DIF and comparison of DIF procedures in a performance assessment. doctoral Thesis, Ehwa Womans University, Seoul.

Stocking, M. L. & Lord, F. M.(1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201~-210.

Stout, W.(1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589~617.

Swaminathan, H. & Rogers, H. J.(1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361~370.

Williams, R. L.(1971). Abuse and misuse of testing black children. The Counseling Psychologists, 2, 62~73

• Received : 09 March, 2017
• Revised : 04 April, 2017
• Accepted : 17 April, 2017